心理科学进展 2023, Vol. 31, No. 6, 887–904
Advances in Psychological Science

https://doi.org/10.3724/SP.J.1042.2023.00887

·研究方法(Research Method)·

基于词嵌入技术的心理学研究：方法及应用*

包寒吴霜[1,2,3] 王梓西[1,2] 程 曦[1,2] 苏 展[1,2]
杨 盈[1,2] 张光耀[1,2,4] 王 博[5] 蔡华俭[1,2]

([1]中国科学院心理研究所行为科学重点实验室，北京 100101)
([2]中国科学院大学心理学系，北京 100049)
([3]英国曼彻斯特大学曼彻斯特中国研究院，曼彻斯特 M13 9PL)
([4]北京师范大学认知神经科学与学习国家重点实验室和 IDG/麦戈文脑科学研究院，北京 100875)
([5]天津大学智能与计算学部，天津 300350)

摘 要 词嵌入是自然语言处理的一项基础技术。其核心理念是根据大规模语料中词语和上下文的联系，使用神经网络等机器学习算法自动提取有限维度的语义特征，将每个词表示为一个低维稠密的数值向量(词向量)，以用于后续分析。心理学研究中，词向量及其衍生的各种语义联系指标可用于探究人类的语义加工、认知判断、发散思维、社会偏见与刻板印象、社会与文化心理变迁等各类问题。未来，基于词嵌入技术的心理学研究需要区分心理的内隐和外显成分，深化拓展动态词向量和大型预训练语言模型(如 GPT、BERT)的应用，并在时间和空间维度建立细粒度词向量数据库，更多开展基于词嵌入的社会变迁和跨文化研究。我们为心理学专门开发的 R 语言工具包 PsychWordVec 可以帮助研究者利用词嵌入技术开展心理学研究。

关键词 自然语言处理，词嵌入，词向量，语义表征，语义关联，词嵌入联系测验
分类号 B841; B849:C91

语言是人类文明的信息化载体。几千年来，人类在文明演进过程中积累了海量语言文本，其中蕴含着大量人类心理和行为信息。然而，直到计算机技术高度发达的 21 世纪，人们才开始以量化的方式利用语言文本探究人类社会、心理和行为规律(Chen et al., 2021; Jackson et al., 2022; Lazer et al., 2009, 2020)。早期的相关研究主要利用语言文本数据中的词频(word frequency)等信息考察一些相对浅层的心理规律(比如个人主义−集体主义水平的变化)。近年来，随着自然语言处理(natural language processing, NLP)技术的发展和成熟(Hirschberg & Manning, 2015)，越来越多的研究开始探讨蕴藏在人类语言中的大量深层次的社会、心理和行为规律(比如个人主义−集体主义文化心理含义的变化)。在自然语言处理的诸多技术中，词嵌入(word embedding)是目前发展较成熟、应用较广泛的一项基础技术，也是各种大型预训练语言模型(pre-trained language model, PLM)的基石。自社会科学领域首个应用词嵌入技术的开拓性研究在 *Science* 发表以来(Caliskan et al., 2017)，其在心理学领域的应用如雨后春笋，目前仍处于爆发式增长中。本文拟全面整理使用词嵌入技术的心理学研究，在厘清现状的同时，展示词嵌入作为一种前沿的心理学研究方法的应用潜力、未来发展方向和需要解决的问题。在梳理现有研究之前，我们首先介绍这些研究的共同基础：词嵌入技术。

收稿日期：2022-08-23
* 国家社会科学基金重大项目“中国社会变迁过程中的文化与心理变化”(17ZDA324)、中国科学院心理研究所自主部署项目“文化变迁与社会适应：行为和影像学的研究”(E2CX3935CX)
通信作者：蔡华俭, E-mail: caihj@psych.ac.cn

1 词嵌入技术：语义向量化表征和语义关联测量[1]

作为自然语言处理的一项基础技术，词嵌入可以量化表示自然语言中词汇的语义，即通过特定算法对语义进行向量化表征，获得词向量(word vector)，从而为后续的智能化语言处理和分析提供基础。基于词嵌入对语义的向量化表征，研究者可以进一步对语义共性和差异进行向量化表征，以及计算不同词语或概念之间的语义关联程度。下面，我们将围绕这三个方面介绍词嵌入技术。

1.1 从“词语”到“向量”：对语义的向量化表征

当我们遇到一个生词，想知道它的含义，一种方法是通过查词典直接了解词义，另一种方法是通过该词在特定语境中的使用情况(特别是它和上下文的关系)推测词义。目前，基于大规模语言文本，计算机对语言的理解主要基于后一种方法，即通过某个词的语用(词与上下文的关系)表征这个词的含义，这就是所谓的“语用即语义”。不过，计算机能够处理的是向量化的语义表征，即词向量。

词向量的发展经历了从简单到复杂、从静态到动态、从机械到智能的过程。研究者先后提出了三种基于数值向量的词汇表征方式(word representation)：独热表示、分布表示、词嵌入表示。

独热表示(独热编码，one-hot encoding)将词表中的 N 个词依次表示为一个 N 维数值向量，每个词向量只有一个维度的值为 1，剩余为 0。独热表示只能简单区分词语，无法表征语义，而且其高维、稀疏的特点容易导致“维度灾难”。为了克服这些局限，研究者提出了词的分布表示(distributional representation)：一个词的语义很大程度由上下文决定，因此语义相近的词往往具有相似的上下文，这就是分布式语义假设的思想(Harris, 1954; Lenci, 2018)。基于这种思想，分布表示将一个词与上下文其他词的共同出现情况(简称共现，co-occurrence)视为这个词的分布结构(distributional structure)，然后使用统计方法对共现矩阵进行降维，最后得到相对低维、稠密的词向量(表 1)。分布表示有两种具体的降维方法。一种方法是潜在语义分析(Latent Semantic Analysis, LSA)，利用奇异值分解实现共现矩阵降维，每个维度反映词的一种独立的潜在语义特征(Landauer & Dumais, 1997)。另一种方法是基于潜在狄利克雷分配(Latent Dirichlet Allocation, LDA)的主题模型(Topic Model)，利用概率分布和贝叶斯统计提取出文本主题，每个维度反映词在相应主题上的出现概率(Blei et al., 2003; Griffiths et al., 2007)。

然而，独热表示和分布表示在大规模语料中的训练速度和效果都欠佳，并且独热表示无法利用上下文信息，分布表示在利用上下文方面效果也不够理想(车万翔 等, 2021)。为了解决这些问题，受到神经概率语言模型(Neural Probabilistic Language Model; Bengio et al., 2003)的启发，研究者提出了词嵌入表示，这是本文讨论的核心。

所谓“嵌入” (embedding)，是指在尽可能保留原有语义距离的前提下，将复杂语义信息直接映射到低维向量空间。词嵌入的理论基础仍为分布式语义假设(Harris, 1954; Lenci, 2018)，与分布表示(LSA、LDA)的差异在于向量赋值方式：词嵌入的降维是通过机器学习算法(一般采用神经网络算法)训练模型来预测词与上下文的共现关系，进而直接获得低维、稠密的词向量(常见的有 300 维，也可根据需要确定维数)。词嵌入表征的语义并不是具象的语义解释，而是算法从人们的自然语言中自动学习的抽象的语用规律，其维度数值本质上是神经网络模型的权重(weights)或通过模型估计得到的参数(类似于回归系数)，一定程度上模拟了人类加工语言时大脑的激活模式。词嵌入对语义向量化表征的示意图见图 1。

词嵌入向量可分为两类：一类是静态词嵌入(static word embeddings)，将一个词在整个语料库中的所有上下文信息都聚合、压缩到一个向量表示中，得到的是固定的、不随词汇所在特定语境中的上下文变化的词向量，算法模型包括 Word2Vec、GloVe、FastText 等；另一类是动态词嵌入(dynamic word embeddings)，也称上下文相关、语境化的词嵌入(contextualized word embeddings)，根据提供的上下文语境得到每个词在特定语境中的词向量，可通过 ELMo、GPT、BERT 等预训练语言模型生

[1] 大部分情况下，“词嵌入”和“词向量”可以互换使用。不过，“词嵌入”侧重于技术思想，即浓缩语义信息并将其映射到低维向量空间，通常也指词嵌入矩阵；而“词向量”侧重于具体数据，也泛指采用词嵌入以外的方法得到的向量，比如本文 1.1 介绍的独热表示、分布表示。广义的“词嵌入”(token embedding)中，词/标记(token)是基本的语义单元，不仅指单词，还包括字、子词(subword)及其他标记信息。

表 1 常见的词向量生成方法和模型

方法/模型	运算处理对象	向量生成方法	维度数值含义
词分布表示	(小规模语料)	(矩阵降维)	(可解释)
潜在语义分析(LSA)	词–文档共现矩阵	奇异值分解(SVD)	独立的潜在语义特征
主题模型(Topic Model)	词–文档共现矩阵	潜在狄利克雷分配(LDA)	词出现于主题的概率
词嵌入表示[提出年份]	(大规模语料)	(训练预测)	(不可解释)
静态词嵌入模型			(不随上下文变化)
Word2Vec[2013]	词–上下文局部窗口	浅层(单层)神经网络	神经网络隐含层权重
– CBOW 子模型	—	(根据上下文预测中心词)	—
– SG/SGNS 子模型	—	(根据中心词预测上下文)	—
GloVe[2014]	词–上下文共现矩阵	加权最小二乘回归	回归迭代求解的参数
FastText[2016]	字符级 *n*-gram 窗口	浅层(单层)神经网络	神经网络隐含层权重
– CBOW 子模型	—	(根据上下文预测中心词)	—
– SG/SGNS 子模型	—	(根据中心词预测上下文)	—
动态词嵌入(语言)模型			(上下文相关)
ELMo[2018]	字符级文本序列	卷积神经网络+双向语言模型	隐含层输出权重组合
GPT[2018]	词–上文文本序列	深层单向转换解码器	隐含层输出权重组合
BERT[2018]	子词文本序列	深层双向转换编码器	隐含层输出权重组合

注：LSA = latent semantic analysis. SVD = singular value decomposition. LDA = latent Dirichlet allocation. CBOW = continuous bag-of-words. SGNS = skip-gram with negative sampling. GloVe = global vectors. ELMo = embeddings from language models. GPT = generative pre-trained transformer. BERT = bidirectional encoder representations from transformers.
神经网络包括输入层、隐含层、输出层。隐含层一般有多个节点(“神经元”)，每个节点为一个激活函数。静态词嵌入模型一般取神经网络前半部分，即隐含层的输入权重矩阵(input weight matrix)，作为词向量矩阵；动态词嵌入模型更复杂，每个词的动态词向量是对该词上下文语义组合的结果，可来自最后一层隐含层的输出权重或多层隐含层向量的加权平均，其中，接近输入层和输出层的隐含层分别编码了更多语法和语义信息(车万翔 等, 2021)。

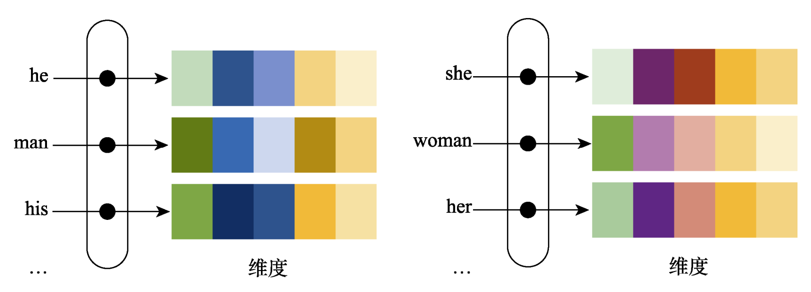


图 1 词嵌入对语义的向量化表征：简化示意图

成(车万翔 等, 2021; 陈萌 等, 2021; Lake & Murphy, 2021)。表 1 总结了这些模型的特点(更多介绍详见网络版附录的补充材料 1)。研究者一般使用在大规模语料上得到了预训练、可表征通用语义知识的词向量(见网络版附表 S1)。

1.2 表征语义共性和差异：词向量的线性运算

由词嵌入技术训练得到的词向量浓缩了词在上下文中的语用规律，一定程度上可以表征人类自然语言中的语义知识。然而，词向量数值的含义通常是不明确的，我们无法直接从向量数值中获得可解释的语义知识。为了获得更明确的含义，一种常见做法是计算语义共性或语义差异的向量表示，即多个词向量相加后的向量总和(vector sum)或相减后的向量差异(vector difference)。[2]

基于词向量的线性运算，我们可以得到词语差异之间的类比(word analogy)，进而获得可解释的语义知识(Mikolov et al., 2013)。比如，语义差异类比可以体现性别差异($\overrightarrow{\text{king}} - \overrightarrow{\text{queen}} \approx \overrightarrow{\text{man}} -$

[2] 只有同一个语义空间(来自相同语料库和模型)的词向量才能进行线性运算，并且需要进行归一化(normalization)，即缩放每个词向量至单位长度 1。

$\overrightarrow{woman}$), 语法差异类比可以体现时态变化($\overrightarrow{work} - \overrightarrow{working} \approx \overrightarrow{play} - \overrightarrow{playing}$), 从属关系类比可以体现国家与首都的关系($\overrightarrow{China} - \overrightarrow{Beijing} \approx \overrightarrow{Japan} - \overrightarrow{Tokyo}$)。

心理学研究中, 词向量的线性运算还有更一般的用途。比如, 研究者可以通过计算与某个心理概念或维度有关的近、反义词的词向量之差, 建立这个概念维度两极的坐标系, 从而使得计算其他心理概念与这个概念间的语义联系成为可能(Kozlowski et al., 2019); 研究者还可以通过计算一系列词向量的总和, 获得这些词的语义共性, 以此表征其上位心理概念。

1.3 测量语义关联和距离: 词向量的联系强度

人类心理的表征在“头脑内”很多时候表现为概念与概念间的联系, 而在“头脑外”的自然语言中则表现为词与词之间的语义联系。因此, 利用自然语言中词与词之间的联系, 我们能在一定程度上探究人类心理特征。

总体上, 语义联系有绝对和相对之分, 计算方法主要有三种: 直接计算词向量的绝对余弦相似度或距离、通过计算两组词向量间的余弦相似度之差获得相对语义相似度(统称为“词嵌入联系测验”)、通过计算两组词向量间的欧式距离之差获得相对语义距离(统称为“相对范数距离”)。下面分别介绍每种方法及其优缺点和适用范围。

1.3.1 余弦相似度和距离[3]

两个词向量在空间中夹角的余弦值, 即余弦相似度(cosine similarity), 可以衡量两个词语之间的语义关联性(semantic relatedness), 其本质上是这两个词的语用或上下文特征的相似性(Lenci, 2018)。余弦相似度取值范围是−1~1, 但一般很少有负数; 与之相反的是余弦距离(cosine distance; 1 − 余弦相似度), 取值范围是 0~2。若两个词完全相关, 则向量夹角为 0°, 余弦相似度为 1, 余弦距离为 0; 若两个词完全无关, 则向量夹角为 90°, 余弦相似度为 0, 余弦距离为 1。

余弦相似度绝对大小的意义并不总是明确。一般而言, 近义词的相似度往往较高, 但相似度高的未必是近义词, 也可能是反义词(如“喜欢−讨厌”)、固定搭配(如“单反−相机”)、语境相近的词(如“键盘−鼠标”)等; 同理, 相似度低的也未必是反义词, 而只是两个毫无关联的词(如“心理−竣工”)。可见, 词相似度衡量的是语义联系的绝对值, 既不必然表示联系的正、负方向, 也难以直接区分同义词和反义词。所以在实际应用中, 为了使词相似度获得可比较的参照点, 研究者一般计算的是相对的词相似度(或距离)。

1.3.2 词嵌入联系测验(WEAT)[4]

心理学研究中, 为了运用概念间的相对语义联系来衡量人们的心理特征, 研究者需要选择能代表特定人群的语料库和由此训练的词向量, 然后计算词的相对余弦相似度。这种分析方法强调目标概念和属性两极之间的相对语义联系, 因此后来被统称为“词嵌入联系测验”。[5]

词嵌入联系测验(Word Embedding Association Test, WEAT)由 Caliskan 等(2017)首次提出, 与内隐联系测验(Implicit Association Test, IAT)的原理和算法类似, 但结论适用范围不同。IAT 为了测量个体头脑中的概念联系, 使用快速按键分类任务测量被试的反应时, 然后将目标概念词(如花−虫)和属性词(如积极−消极)在不相容和相容条件下的反应时之差作为态度、偏见、刻板印象等心理

---

[3] 余弦相似度的计算公式:

$$\text{similarity}(A,B)=\cos(\vec{A},\vec{B})=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}A_i^2}\times\sqrt{\sum_{i=1}^{n}B_i^2}}$$

当两个向量的模长经过归一化, 它们的欧氏距离与余弦相似度或距离存在固定关系: $\|\vec{A}-\vec{B}\|=\sqrt{2(1-\cos(\vec{A},\vec{B}))}$

[4] WEAT 计算两类目标概念(如 X=花, Y=虫)和两类属性词(如 A=积极, B=消极)的相对相似度。首先计算一组目标词(X 或 Y)中的某个具体词 w 与属性 A 和 B 的词相似度之差, 作为 w 与属性两极的相对相似度; 然后计算目标概念 X 和 Y 与该属性相似度的差值, 作为目标和属性间的相对联系强度。WEAT 的计算公式:

$$s(w,A,B)=\text{mean}_{a\in A}\cos(\vec{w},\vec{a})-\text{mean}_{b\in B}\cos(\vec{w},\vec{b})$$
$$s(X,Y,A,B)=\text{mean}_{x\in X}s(x,A,B)-\text{mean}_{y\in Y}s(y,A,B)$$

单类 SC-WEAT 则是计算单类目标概念的每个词和两类属性词的相似度均值差异, 即上述公式的第一步。

[5] WEAT 也被译为“词嵌入联想测验” (吴胜涛 等, 2020)。Association 本身有“联系”和“联想”两种含义。本文建议此处译为“联系”, 原因如下: (1)起初 IAT 被译为“内隐联想测验”, 但原始译者和多位学者已将 Association 的翻译修正为“联系” (杨紫嫣 等, 2015); (2)严格来说, WEAT 并不测量个体头脑内的联想加工过程, 而是测量自然语言中不同词语之间的语义联系(Caliskan et al., 2017)。但与 WEAT 不同, 本文 2.3.1 节介绍的 Divergent Association Task 中的 Association 译为“联想”更合适, 因为该测验涉及个体在任务中的自由发散联想过程(Olson et al., 2021)。

特征的间接测量指标(Greenwald et al., 1998; 杨紫嫣 等, 2015)。WEAT 则是将词相似度视为 IAT 中的反应速度, 使用两组目标词和两组属性词的词相似度之差来测量目标概念词和属性词在自然语言中的相对联系强度, 并且可以使用 d 值衡量标准化效应量(Caliskan et al., 2017)。

同时, 为了考察单类目标概念(如职业)与两极属性(如性别)的联系, Caliskan 等(2017)还提出了词嵌入事实联系测验(Word Embedding Factual Association Test, WEFAT), 后来被称为单类 WEAT (single-category WEAT, SC-WEAT; Toney-Wails & Caliskan, 2021)。这种单类 WEAT 和单类 IAT (SC-IAT; Karpinski & Steinman, 2006)类似, 允许研究者只考察单个目标概念而无需找到与之相对的另一个目标概念, 但属性词仍要有两极对比。

WEAT 和 SC-WEAT 是目前在心理学研究中应用最多的基于词向量的概念相对联系测量方法。但是, 在使用群体大规模语料的前提下, 其测量的并不是个体头脑中的概念联系, 而是概念联系在特定时空下的某个语料库中的一种外化表达, 由此仅能推测产生语料的相应群体的心理特点。因此, 虽然 WEAT 和 IAT 的结果可以做类似理解, 比如都能用来测量刻板印象、偏见等, 但 WEAT 反映的是群体水平的概念联系, 而 IAT 测量的则是个体水平的概念联系。

1.3.3　相对范数(欧氏)距离(RND)[6]

概念间的相对联系还可以通过 Garg 等(2018)提出的相对范数距离(relative norm distance, RND)来衡量。相对范数距离又称相对欧氏距离(relative Euclidean distance; Bhatia & Bhatia, 2021), 计算的是一个目标词(比如某职业)和两类属性词(比如男性和女性)的词向量欧式距离之差。

RND 与 SC-WEAT 类似, 都用于衡量单类别目标概念与一对属性的相对联系, 只是解释的方向相反。SC-WEAT 数值表示相对语义相似度, 因此数值越大表示概念相对联系越紧密; 而 RND 数值表示相对语义距离, 因此数值越小表示概念相对联系越紧密。二者算法上的区别对结果的实际影响不大, 研究者可根据实际需求选用其中一种指标。

总之, 根据词嵌入技术对词汇语义的表征, 研究者可以较好地量化文本中的语义信息。具体地, 研究者不仅能通过词向量的线性运算获得语义共性或差异的表征, 而且能通过计算余弦相似度、欧氏距离、基于余弦相似度的 WEAT 与 SC-WEAT、基于欧氏距离的 RND 等方法衡量概念间的语义联系。利用这些前沿技术和方法, 研究者就能通过自然语言来量化并探究人类的心理和行为规律。

2　基于词嵌入技术的心理学研究

自从 Mikolov 等(2013)首次提出词嵌入算法, 特别是 Caliskan 等(2017)首次将词嵌入技术应用于社会科学领域以来, 基于词嵌入的心理学研究在短短几年内大量涌现, 内容涉及心理语言学、决策判断、心理健康、社会认知、人格心理、道德心理、政治心理、文化心理等众多心理学领域。而与这些研究有关的一项基础工作是利用词向量相似度来辅助构建合理、有效的心理概念词表。因此, 下面将首先介绍词嵌入在词表构建方面的应用。

2.1　构建心理词表: 研究的基础工作

词向量首先可以用来构建心理概念词表, 包括使用词相似度评估词表的信效度、扩充近义词等。例如, 一项研究在构建刻板印象内容(stereotype content)词表时, 使用词向量计算了每两个词的相似度, 发现同一维度内的词相似度高于不同维度间的词相似度, 以此验证了词表的内部一致性信度和区分效度(Nicolas et al., 2021)。还有研究者借助词相似度为初步构建的词表扩充近义词, 并结合专家评估进一步确定复杂概念的词表, 比如文化松紧性(tightness−looseness; Jackson et al., 2019)、本真性(authenticity; Le et al., 2021)等。此外, 基于预训练语言模型, 清华大学的研究团队开发了

[6] 范数(norm)是线性代数的术语, 表示向量在空间中的长度(模长), 此处指 L2 范数(欧式范数)。两个向量的范数距离表示它们差异的长度, 衡量了它们在空间中的距离; 词向量的范数距离衡量了语义距离。由此, 相对范数距离(RND)计算目标概念和两类属性词的相对语义距离。比如, 对于职业(目标概念)和性别(属性)的联系, 首先分别计算男性和女性所有词的平均向量, 然后计算每个职业的词向量与这两个平均向量的欧氏距离之差, 即单个职业的 RND, 最后计算所有职业 RND 之和。结果若为负值(职业和男性词向量的欧式距离小于和女性词向量的欧式距离), 则说明职业与男性的联系比女性更紧密。RND 的计算公式:

$$\text{relative norm distance}(W, A, B) = \sum_{w \in W} \left(\left\| \vec{w} - \text{mean}_{a \in A}\, \vec{a} \right\| - \left\| \vec{w} - \text{mean}_{b \in B}\, \vec{b} \right\| \right)$$

WantWords 反向词典平台(https://wantwords.net)，可以帮助研究者根据定义、词性、字数、包含的字词等方面精准查找近义词。总之，利用词相似度辅助构建词表，能避免人工选词过程中可能存在的主观偏差，提高词表的规模、信度、效度和代表性，最终增强研究的客观性和可靠性。

接下来，我们以“语义”作为逻辑线索，将词嵌入技术在心理学研究中的具体应用分两类介绍：基于语义表征的研究和基于语义关联的研究。

2.2 基于语义表征的研究

2.2.1 帮助探究人类语义加工的脑活动

词向量作为计算机对语义的向量化表征，能用来帮助考察人类语义加工的脑活动。具体来说，在使用神经影像测量仪器(如功能性磁共振 fMRI)记录被试大脑活动的基础上，研究者可以构建词向量与相应的词诱发的大脑神经活动之间的映射关系模型(词的神经响应模型)，进而预测大脑对其他词汇语义和语义关系加工的特异性脑活动。例如，一项发表于 *Nature* 的研究使用 fMRI 记录被试听故事时的脑活动，在分析时对实验材料的每个词分别构建其与一系列基础词汇在既有语料库中的共现频次，以此作为词向量，进而构建每个故事在这些维度上的语义向量的时间序列矩阵；然后利用机器学习，发现基于这种语义向量构建的神经响应模型能有效预测额叶、颞叶等脑区的激活，说明这两个脑区在语义表征中具有重要作用(Huth et al., 2016)。与之方法类似，另一项研究使用每对词的词向量之差表示其语义关系(如“手–手指”反映了从整体到部分的语义关系)，结果发现基于这种语义关系向量构建的机器学习模型也能预测特定脑区的激活(Zhang et al., 2020)。

2.2.2 预测人们对特定事物的认知判断

词向量在探究人类认知加工方面的应用还可以拓展到更复杂的形式。基于词向量原始值构建的机器学习模型能预测人们的各类认知判断结果，从而可以对人类的复杂认知判断进行更准确的计算建模(Bhatia et al., 2019)。研究者将预训练的词向量原始值作为预测变量(每个维度是一个变量)，将人们对不同事物的评价作为结果变量(单个事物获得的多人评价平均值)，使用岭回归(ridge regression)等算法构建机器学习模型，进而预测人们对公众人物和其他事物的认知评价。例如，一项研究通过建立岭回归模型，把公众人物名字的词向量作为预测变量，把这些人物被人们评价的领导力作为结果变量，发现模型可以根据人名词向量预测人们感知到的领导力(Bhatia et al., 2022)。基于这类模型，研究者还可以根据事物名词的词向量预测人们对风险源(Bhatia, 2019a)、食品健康程度(Gandhi et al., 2022)、身体健康状态(Aka & Bhatia, 2022)、食物热量和婴儿死亡率(Zou & Bhatia, 2021)的认知判断，以及社会认知、风险感知、健康行为、组织行为和市场营销等领域中的复杂认知判断(Richie et al., 2019)。这些研究都是直接利用原始词向量中的语义信息，并将其用于建立行为预测模型。

此外，研究者基于特征属性两极的向量差异构建语义特征维度(比如大小、安全–危险程度)，将词向量在不同维度上分别进行语义投影(semantic projection)，结果发现，经过语义投影的词向量在相应维度上的位置可以预测人类对这些事物相应属性的判断(Grand et al., 2022)。这种方法仿照心理量表的形式，利用词嵌入对语义差异的表征，不仅实现了对事物属性的自动化评估，而且还原了蕴含在词向量中的丰富的语义信息和人类知识。

2.2.3 评估个体的情绪和心理健康

还有研究将原始词向量作为机器学习模型的输入参数，以此建立预测模型，实现对个体情绪和心理健康状况的评估。例如，研究利用 BERT 模型，将个体静息状态下的自发思维内容(句子)转换为“片段向量”，然后使用有监督的深度学习来训练情绪分类模型，从而识别个体自发思维内容的情绪类型(H.-X. Li et al., 2022)。也有不少研究沿用类似手段，基于微博等社交平台用户自发产生的文本，使用词向量原始值和机器学习模型识别个体的心理健康状况和精神障碍，包括抑郁、焦虑、压力、自杀风险等(Kalyan & Sangeetha, 2020; Salas-Zárate et al., 2022)。同时，在词向量基础上考虑用户的人口学变量和微博行为(王垚 等, 2022)、多模态信息(Lin et al., 2020)等，能进一步提高对心理症状识别的准确性。

2.3 基于语义关联的研究

2.3.1 评估和探究个体心理

利用词向量相似度衡量的语义关联，并借助专门设计的研究范式，研究者可以评估和探究部分个体心理，目前主要涉及发散思维能力、决策倾向等。

首先, 将词向量距离指标与心理测量任务相结合, 研究者可以更客观地评估个体的发散思维(远距离联想)能力。研究者提出了发散联想任务(Divergent Association Task, DAT): 施测时让被试思考并列出 10 个相互尽可能无关的名词; 然后利用预训练的词向量计算这些词两两之间的余弦距离; 这种根据被试列举的若干名词计算的平均语义距离可以反映个体在多大程度上能想出距离较远的事物, 语义距离越大, 则说明个体的发散思维越强(Olson et al., 2021)。类似地, 其他研究者也提出了利用语义距离测量发散思维的方法(Beaty & Johnson, 2021; Heinen & Johnson, 2018; Johnson et al., 2021)。这些结合词向量语义距离的测量方法弥补了创造力传统测量工具的局限: 一方面, 测量无需依赖自评或专家评定, 避免了主观性; 另一方面, 实际施测时只需要请被试自由列举一系列词汇, 并由此计算平均语义距离, 提高了测量的便捷性, 有助于大规模施测。

其次, 词相似度衡量的语义关联能反映个体决策中的联想加工倾向。例如, 研究者在多种决策情境中比较了问题文本与不同选项文本间的语义相似度, 结果发现人们倾向于选择与题干语义最相近的选项(Bhatia, 2017a)。使用类似方法, 研究者还验证了决策的语义聚集效应, 即个体在选择情境逐一给出回答时, 倾向于搜索与已经想到的回答语义相近的答案(Bhatia, 2019b)。因此, 词向量蕴含的语义关联信息有助于研究者更准确地探究个体的决策倾向与选择偏好。

2.3.2　评估和探究社会心理

现有的词向量通常是由一个群体产生的大规模文本语料训练出来的, 这些文本可能蕴含群体的心理特征。因此, 基于词向量(目前主要是静态词向量)计算的语义关联, 包括 WEAT、RND 等概念相对联系指标(见 1.3 节), 可专门用于测量群体的心理特征, 比如群体的社会态度、刻板印象、社会偏见、道德偏差、文化心理联系等, 以及上述心理现象的产生、发展和演变。

在 WEAT 提出之前, Bolukbasi 等(2016)发现性别词向量之差(如“she−he”)与职业词向量之差(如“nurse−surgeon”)的余弦相似度能预测人工评价的性别−职业刻板印象。受其启发, Caliskan 等(2017)发表在 *Science* 的研究进一步提出了 WEAT 和 SC-WEAT, 用来测量群体的社会认知, 并重复了内隐社会认知领域的多项经典结果, 包括花−虫内隐态度、乐器−武器内隐态度、内隐种族偏见、内隐性别−职业刻板印象、内隐性别−学科刻板印象等。这两项奠基性研究迅速激发了一系列研究直接应用 WEAT 或类似方法测量各类社会认知, 例如: 对不同颜色的态度和性别−颜色刻板印象(Jonauskaite et al., 2021)、对不同职业和国籍群体的人格特质刻板印象(Agarwal et al., 2019)、不同语言中的性别偏见(Kurpicz-Briki & Leoni, 2021)、法律文书中的种族偏见(Rice et al., 2019)、新闻报纸中的种族偏见和性别刻板印象(Bhatia, 2017b)、电影和文学作品中的性别刻板印象(Xu et al., 2019)、人类集体概念(collective concept; PERSON/PEOPLE)的性别偏差(Bailey et al., 2022)、群际态度(评价)和群际信念(刻板印象)之间的关系(Kurdi et al., 2019, Study 3)、企业组织语境中的性别−领导力刻板印象及其与女性领导雇佣比例之间的相互影响(Lawson et al., 2022)、不同政治倾向或党派的新闻媒体对政治内群体的积极态度偏差和对政治外群体的消极态度偏差(Rozado & al-Gharbi, 2022)等。

同时, WEAT 和 SC-WEAT 还被用于探究群体的道德偏差。一项研究使用 WEAT 考察了“自我−他人”目标词和“道义主义−功利主义”属性词的相对语义联系, 结果发现自我(vs.他人)与道义(vs.功利)的联系更紧密, 说明人们倾向于认为别人是功利而非道义的, 揭示了群体层面的自我−他人道德偏差(M.-H. Li et al., 2021, Study 3)。另一项研究则使用 SC-WEAT 考察了“正义”单类目标词和“自我−他人”属性词的相对语义联系, 结果发现正义与他人(vs.自我)的联系更紧密, 据此推测正义动机可能存在他人凸显效应(吴胜涛 等, 2020)。

此外, 类似方法还能用于分析公众人物被人们感知到的人格特质。一项研究基于公开新闻语料计算了美国前总统候选人唐纳德·特朗普(Donald Trump)和希拉里·克林顿(Hillary Clinton)的人名向量与各种人格特质评价(如温暖、能力、道德)词向量的相对余弦相似度, 以此衡量大众感知到的二人的人格特质(Bhatia et al., 2018)。这类研究可以在非接触条件下, 间接测量人们对公众人物的人格特质的感知, 从而弥补传统量表工具难以用于公众人物的局限, 也有助于探究与政治人物有关的问题。

除了使用 WEAT 等方法直接测量群体层面的社会心理特征，还有不少研究进一步探究了刻板印象、偏见等社会心理现象的产生、发展和变迁。

首先，关于社会认知的产生，目前有两项研究采用 WEAT 测量了多个国家的社会刻板印象或偏见，发现语言特征可能会塑造和加深人们的社会认知。其中，一项研究选取 25 种语言，发现语言中的性别−职业刻板印象 WEAT 分数和性别化职业词(如 waiter/waitress)的比例均能正向预测国家层面的内隐性别−职业刻板印象 IAT 结果，说明语言可能会塑造内隐社会认知(Lewis & Lupyan, 2020)。另一项研究则将 45 种语言分为性别化语言(gendered language; 名词、动词和形容词有阴阳性之分，如法语、西班牙语)和无性别语言(genderless language; 词语不区分阴阳性，如汉语、英语、芬兰语)，计算了每种语言的性别偏见 WEAT 分数，结果在性别化语言中发现了更大的性别偏见，说明一门语言的语法规则可能会加深社会偏见(DeFranza et al., 2020)。这些研究利用词嵌入的方法优势和多语种词向量库的丰富资源，巧妙解决了此前难以直接回答的理论问题。

其次，关于社会认知的发展，目前也有两项研究采用 WEAT 测量并追溯了性别刻板印象在儿童发展早期的表现。其中，一项研究收集儿童和成人语料库并训练词向量，计算了性别刻板印象的 WEAT 分数，结果发现性别刻板印象存在于不同年龄的语言中(Charlesworth et al., 2021)。另一项研究则使用亲子对话语料库，计算了词汇被不同性别使用的概率、词汇−性别的概念联系 WEAT 分数及两者相关，结果发现 2~5 岁儿童已经有了性别化的语言表达(Prystawski et al., 2020)。这些研究同样利用词嵌入的方法优势，巧妙实现了对婴幼儿群体的心理测量。

最后，基于语义关联的历时性演变，不少研究利用词向量探讨了社会认知与文化心理的变迁。社会与文化变迁是近年来心理学、社会学的前沿研究热点(蔡华俭 等, 2020; 黄梓航 等, 2018, 2021)。以往研究主要是利用调查数据、历史档案数据、过去发表的研究数据等考察某个心理现象的均值或水平的变迁，而较少能探讨概念含义或概念之间关系的变迁(蔡华俭 等, 2023)。利用跨时间的词向量库，为每个年代或年份分别计算语义联系指标并形成时间序列，可以考察社会态度、偏见、刻板印象、概念的文化含义、文化与心理的关系等方面的变迁。

现有研究主要使用了 HistWords 项目预训练好的以十年为单位的词向量库(Hamilton et al., 2016)，然后为每个年代分别计算语义关联指标(如 WEAT 或 RND)，分析刻板印象与偏见的变化；或提取出每个年代与目标概念(如社会群体)联系最紧密的特质词，并分析这些词的效价(积极/消极)等属性的变化。基于此，研究者揭示了：美国社会的性别刻板印象和种族刻板印象在 20 世纪逐渐减弱(Bhatia & Bhatia, 2021; Garg et al., 2018); 社会的不同属性维度(如贫−富、男性化−女性化、道德高低、教养高低等)及不同维度之间的关系在 20 世纪的变化(Kozlowski et al., 2019); 新闻媒体对种族外群体的刻板印象内容从 2005 到 2015 年的变化(Kroon et al., 2021); 人们对 14 类社会群体(包括不同性别、种族、年龄、体型和社会阶层的群体)的刻板印象内容及其效价从 1800 到 2000 年的变化(Charlesworth et al., 2022); 道德概念、道德的积极−消极效价和道德基础维度(如关爱−伤害、公平−欺骗)从 1800 到 2000 年的变化(Xie et al., 2019)。此外，一项研究利用谷歌图书和《纽约时报》语料库，分别使用词频分析、情感分析、主题模型分析和词嵌入分析，揭示了 1800~2000 年风险(risk)概念的词频在上升，情感效价越来越消极，主题从战争转向疾病，语义逐渐趋近于对风险的规避和预防(Y. Li et al., 2020)。而关于文化心理变迁，Hamamura 等(2021)考察了中国的个人主义/集体主义与其他 10 个概念(如积极、消极、成就、金钱、休闲、工作、家庭等)之间的联系从 1950 到 2000 年的变化；根据对其结果的重新分析和正确解读，个人主义越来越被中国人接受(态度从消极变为中性)，并且与富裕(而非贫穷)、休闲娱乐等方面的联系变得更紧密(Bao et al., 2022)。

2.4　小结

总之，由大规模语料训练出来的词向量不仅表征了社会文化中的语义信息，而且蕴含了许多人类心理和行为信息。心理学研究可以利用词向量的原始值(向量)、线性运算结果、绝对相似度或距离、相对相似度或距离，考察蕴含在词向量或其关系背后的心理和行为现象及其规律。表 2 总结了这些应用形式、用途特点和利用的语义信息。

表 2　词向量在心理学研究中的应用形式、用途特点和利用的语义信息

应用形式	具体分析指标	用途特点	利用的语义信息
词向量			
原始值	数值向量	作为机器学习模型(比如岭回归)的输入参数，预测个体的认知判断、大脑活动等	语义表征
线性运算结果	相加后的向量总和、相减后的向量差异	作为语义共性或语义差异的量化表示，或以向量差异建立一个心理概念维度两极的坐标系，计算与其他心理概念间的语义联系	语义共性和差异的表征
词向量的关系			
绝对相似度	余弦相似度、欧氏距离	作为词汇或概念间语义联系的直接测量指标，测量个体的发散思维水平(远距离联想)等	语义关联或距离
相对相似度	词嵌入联系测验(WEAT)、单类词嵌入联系测验(SC-WEAT)、相对范数(欧氏)距离(RND)	作为词汇或概念间语义联系的相对测量指标，测量复杂概念间的联系，如群体的社会态度、偏见、刻板印象、文化与心理的联系等	语义关联或距离

3　讨论

现代科学心理学始于 1879 年冯特在德国莱比锡大学建立的第一个心理学实验室。大家所熟知的是，通过建立第一个心理学实验室，冯特为科学研究人类心理与行为指明了一个基本途径，即通过直接观测和分析人的心理与行为来研究其规律；然而不太为大家所知的是，冯特晚年专注的民族心理学其实还为研究人类心理与行为指明了另外一种途径，即研究包含大量人类心理与行为信息的各种产品。一百多年来，心理学的绝大多数研究都是基于冯特开创的第一个途径开展的。近年来，随着计算机、人工智能和自然语言处理技术的突飞猛进，通过文化产品和自然语言来探索人类心理和行为规律的研究开始涌现。作为自然语言处理的关键技术，词嵌入近年来在心理学研究中得到了越来越多的应用。为了促进词嵌入在中国心理学界的普及和应用，本文对词嵌入的基本方法及其在心理学领域的各种应用进行了至今最全面的介绍。下面的讨论中，我们将首先总结该方法在心理学中应用的基本流程，然后分析其优缺点和主要问题，最后试图指明重要的未来研究方向。

3.1　运用词嵌入方法开展心理学研究的基本流程

为了便于大家更好地掌握词嵌入方法在心理学研究中的应用流程，根据前面两部分的介绍和整理，我们构建了一个基于词嵌入的心理学研究的整体框架(图 2)。从图 2 可以看出，总体上，基于词嵌入的心理学研究通常是数据和理论共同驱动的。数据驱动部分的词向量训练为研究提供必需的语义特征向量，理论驱动部分的问题提出和假设推导则为词向量的应用指明方向。在词表构建过程中，数据和理论都不可或缺。有了合理的词表和预训练好的词向量，研究者就可以根据研究目的，选取恰当的词向量分析指标来开展心理学研究，包括对心理和行为的描述和预测。

3.2　词嵌入方法的优势

与传统的对人的心理和行为直接观测和分析的方法相比，词嵌入方法具有多方面的独特优势。

第一，研究成本低。使用词嵌入方法几乎不需要考虑招募被试的成本；同时，如果使用现成的预训练好的词向量库，则只需要一台普通的计算机即可完成分析。而传统的行为实验、问卷、访谈等都需要人工招募被试，研究周期较长，被试费成本较高。

第二，样本代表性高。词向量通常是根据大规模文本语料训练的(比如 Common Crawl 语料库覆盖了多种来源、万亿级规模的网页链接)，分析结果更能代表人群总体。而传统方法中，样本量一般比较有限，且以学生样本居多，只有经过严格、系统的抽样才能保证样本代表性。

第三，分析客观性强。词向量是通过机器学习算法自动训练而来的，全程少有人为干预，虽然语言本身是由人类产生的，但对语言的分析是量化、自动化、无需依赖人类主观报告的，因此分析过程具有相对客观性。而传统基于被试自我报告的方法容易受到主观性、社会赞许性和反应偏差的影响。

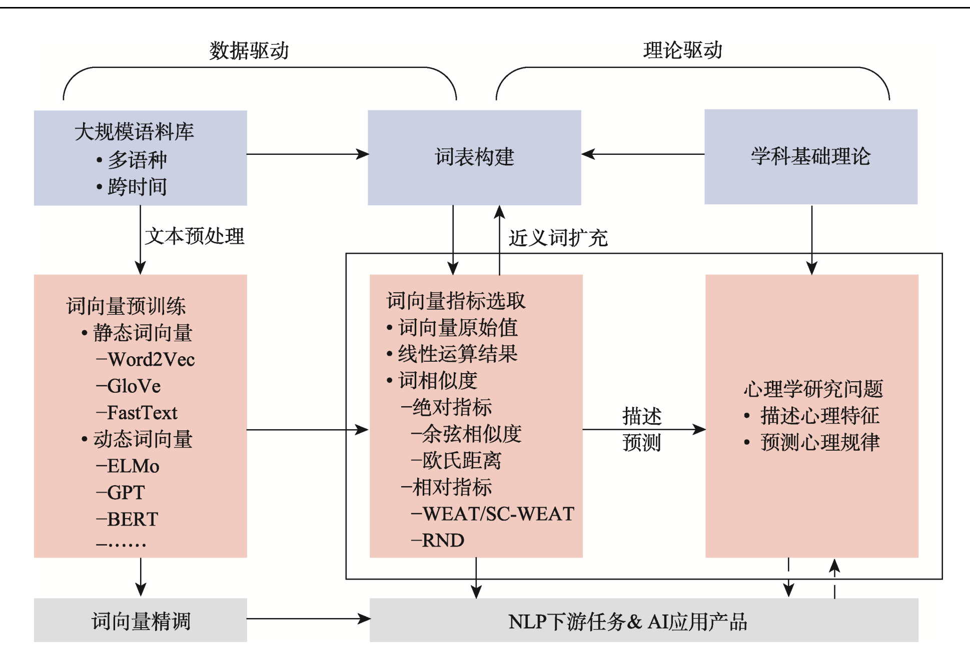


图 2 基于词嵌入技术的心理学研究：整体框架

第四，研究结果可重复。如果研究者使用相同的语料库、预训练词向量库、词典和分析方法，则理论上可以获得完全一致的结果。因此，在当前社会科学面临可重复性危机的背景下，词嵌入方法具有明显优势。

第五，研究主题灵活。研究者借助词向量库可以分析任意词语或概念间的语义联系，从而可以灵活选取研究主题。如果要纳入更多的词，则不需要重新收集数据，只需要增加相应的词。

第六，特别适用于研究某些特定问题。虽然对于一些研究主题(如语义加工脑活动)，词嵌入仅起到辅助作用，但如果研究者想要探究横跨数百年的社会认知与文化变迁，或考察几十种语言文化中的社会认知，或大规模快速测量个体的发散思维能力等，则词嵌入是不可或缺的方法。

3.3 词嵌入方法的局限

尽管词嵌入有诸多优势和独特价值，心理学研究者仍需认识到其局限。

首先，计算机算法学习语义的过程只是利用词共现信息估计模型参数，与人类习得语言的复杂过程相差甚远，因此词向量难以对语义背后深层的心理机制(如动机、目标、意图等)进行表征(Lake & Murphy, 2021)，也难以直接反映人们头脑内的主观心理过程。

其次，在理解词向量时，语料及其产生群体的特点和性质是非常关键的，因为这在很大程度上决定了词向量的意义和结论的适用范围。比如，基于新闻报道训练的词向量反映了媒体记者的语用特征和心理特点，而基于微博训练的词向量反映了微博用户的语用特征和心理特点。词向量只能用于探究对应群体的心理，不能推广到其他群体。因此，在心理学研究中选取词向量数据时，要尽量保证语料库来源与研究问题相符和匹配，否则得到的结论未必正确。

再次，WEAT、SC-WEAT、RND 等方法一般是基于群体语料及由此训练的词向量，在此情况下只能测量群体的社会认知，不能像 IAT 一样测量个体的社会认知(Caliskan et al., 2017)。从某种意义上讲，WEAT 等前沿方法和 IAT 等传统方法是互补的，必要的话可以结合起来使用(e.g., Kurdi et al., 2019; M.-H. Li et al., 2021; Rheault & Cochrane, 2020; Rozado & al-Gharbi, 2022)。

最后，目前基于词向量的心理测量在信度(Du et al., 2021; Durrheim et al., 2023; Richie & Bhatia, 2021)和效度(Joseph & Morgan, 2020; Rodman, 2020)方面仍存在一定争议。为了增强结果的稳健性和结论的说服力，研究者需要构建尽可能充足、全面的近义词表，而不应只依赖少量关键词。同时，对于同一个研究问题，研究者可以将词嵌入方法和传统文本分析方法(如词频分析、

主题模型分析)结合起来, 从而充分挖掘文本中蕴含的心理规律(Arseniev-Koehler et al., 2022; Y. Li et al., 2020)。

3.4　词嵌入心理学研究的重要问题

虽然词嵌入方法在心理学研究中的应用发展迅猛, 但是依然存在一些重要的基础性问题。下面我们对其中关注度比较高的三个问题进行分析和讨论。

3.4.1　如何有效解释词向量维度?

在大部分词嵌入算法中, 词向量的维度本质上是神经网络模型的隐含层权重或输出权重组合。因此, 词向量对语义的表征是抽象的, 难以从语言学角度解释, 也难以确定每个维度究竟代表哪种语义特征。为了增强词向量的直观性, 研究者往往会使用一种降维算法: t 分布随机近邻嵌入(t-Distributed Stochastic Neighbor Embedding, t-SNE)。词向量常见的几十到几百维对于人类而言仍属于高维信息, 而 t-SNE 算法可以将词向量嵌入到二维或三维空间, 同时尽量保留原始向量空间中的语义距离(Hinton & Salakhutdinov, 2006; van der Maaten & Hinton, 2008)。图 3 举例展示了 t-SNE 降维后的可视化结果。可见, 降至平面的词向量不仅较好地保留了词汇间的语义距离和类比关系, 而且使这些语义关联的解释更直观。

然而, t-SNE 维度仍然是抽象的, 不表示具体语义; 而且 t-SNE 是随机过程, 每次都产生不同结果。如果想从词向量不可解释的维度中提取出可解释的语义信息, 可事先确定语义维度并建立坐标系, 然后计算每个词与维度两极的相对相似度(Kozlowski et al., 2019)或进行语义投影(Grand et al., 2022); 此外, 还可以使用主成分分析、有监督的机器学习等方法(Günther et al., 2019; Utsumi, 2020)。

3.4.2　如何区分不同的心理特征?

词向量是多方面因素共同作用的复杂产物, 因此由词向量或词向量的关系指标得到的结果可能是多种心理特征的混合, 比如情绪词反映的情绪可能是理想情感(ideal affect)和实际情感(actual affect)的混合(Tsai, 2007)、认知偏差可能是外显(explicit)和内隐(implicit)认知的混合(Greenwald et al., 1998)。

具体到词嵌入研究, 虽然自 Caliskan 等(2017)基于IAT的思想提出WEAT和SC-WEAT以来, 大量研究应用这些方法考察了文本中蕴含的社会态度、偏见和刻板印象, 但目前我们仍不清楚 WEAT 测量的社会认知是外显的、内隐的还是二者的混合产物。

为了区分 WEAT 测量中的外显和内隐认知成分, 研究者提出了一种可能的解决思路: 将 WEAT 的目标词分为概念词(如“花”)和范例词(如“玫瑰”、“郁金香”), 而态度属性词保持一致(如积极−消极); 然后将概念词与属性词的 WEAT 分数作为

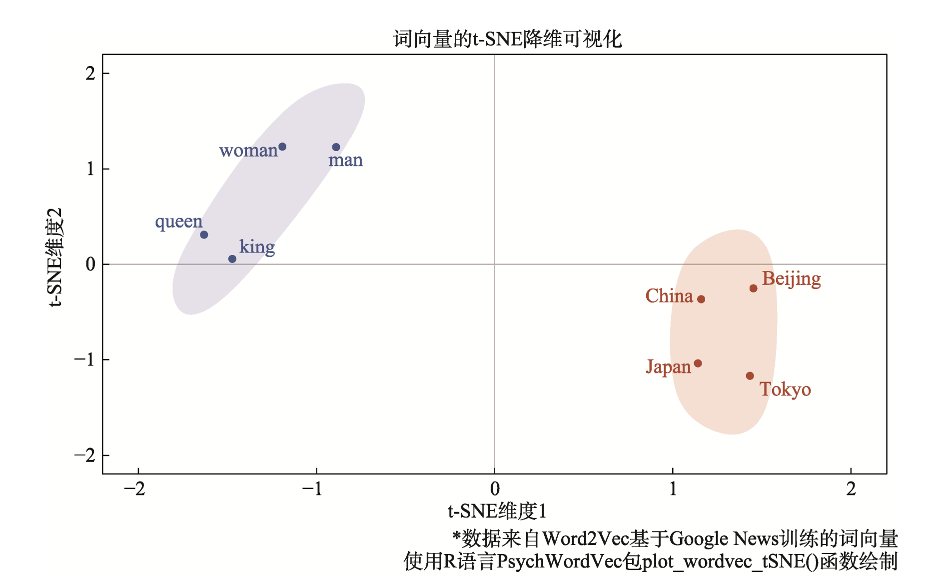


图 3　基于 t-SNE 降维算法的词向量可视化

外显态度，范例词与属性词的 WEAT 分数作为内隐态度(Wang et al., 2019; 薛栢祥, 2019)。不过，该方法只区分了目标词的性质，而且概念词数量可能不足 5 个(甚至只有 1 个)，远小于范例词数量，容易使结果不稳定。本文认为，区分属性词的性质也许是一种更合适的解决办法。比如，对于积极–消极属性，直接描述态度对象的评价性词汇(如“芳香”、“恶臭”)可能反映外显态度，而间接关联的非评价性词汇(如“健康”、“疾病”)可能反映内隐态度。总之，关于如何有效区分词向量概念联系指标的外显和内隐成分，目前尚无充分、直接的实证证据，未来需要深入探讨。

3.4.3 绝对还是相对的词相似度？

词相似度指标既可以是绝对的(如余弦相似度、欧氏距离)，也可以是相对的(如 WEAT、RND)。在研究中，我们应该使用绝对还是相对指标呢？实际上，原始的词相似度很难体现其效应大小，目前并没有关于词相似度大小的明确划定标准；如果没有参照点，则绝对值难以解释。因此，大部分研究都采用相对的词相似度(e.g., Caliskan et al., 2017; Garg et al., 2018; Kozlowski et al., 2019)。特别是，当涉及存在效价或极性的概念(如积极–消极)时，如果不区分而将其混在一起分析，则研究结论可能有偏误(e.g., Hamamura et al., 2021)。

由于词相似度无法区分反义词，研究者需要人工将效价或极性相反的词(反义词)明确区分开(Bao et al., 2022; Grand et al., 2022; Kozlowski et al., 2019; Lee et al., 2021; Nicolas et al., 2021)。例如，道德概念中的积极词(“善”的方面)和消极词(“恶”的方面)构成了道德维度两极，研究者可以将近义词、反义词的词向量配对相减得到差异向量，然后以此为基准计算单个目标词向量与该差异向量的相似度(Izzidien, 2022)。相对指标更少受到共变因素干扰，心理学含义也更明确。例如，一项研究发现大部分词的绝对相似度都存在下降趋势，作者认为一种可能的原因是语言复杂性的增加(Hamamura et al., 2021)；但是，词的相对相似度没有明显的整体下降趋势(Bao et al., 2022)。

3.5 词嵌入心理学研究的未来展望

词嵌入作为一项新兴的自然语言处理技术，在心理学研究中已经并将继续展示巨大的应用潜力。下面我们聚焦三个亟待未来研究的方向。其中，前两个涉及方法，第三个涉及实质性的心理学研究。

3.5.1 建立细粒度的词向量数据库[7]

虽然目前已有的词向量库已经初步划分了年代和语种(见附表 S1)，但为了探究更细水平的心理规律，比如将年代细化至年份，或将语种和国家细化至同一国家内部的不同地区，则需要额外训练获得细粒度(fine-grained)的词向量数据。首先是时间维度。HistWords 词向量库的时间粒度仅到年代(每十年分别训练的词向量)，时间点较少(中文词向量仅覆盖 5 个年代)，难以满足变迁研究的需求，也难以在其中应用时间序列分析方法，如格兰杰因果检验(蔡华俭 等, 2023)。同时，HistWords 项目仅使用了谷歌图书语料，目前暂时缺少基于其他语料的跨时间词向量库。为了克服这些局限，未来研究有必要使用更多来源的语料，如《人民日报》、《新闻联播》、微博等，建立以年为单位的词向量库。自行训练词向量时，可采取一些策略弥补每年文本量的不足，增强年度词向量数据的稳健性：采用三年滑动窗(某年及前后各一年)的全部文本作为该年的文本来训练词向量，相当于从源头进行平滑化(Garg et al., 2018; Lawson et al., 2022)。此外，为了使词向量具有跨时间可比性，即解决不同时期向量空间的对齐问题(alignment problem)，一般可以使用 Schönemann (1966)提出的正交普鲁克(Orthogonal Procrustes)矩阵对齐方法(Y. Li et al., 2020; Hamilton et al., 2016; Rodman, 2020)。

另一个需要细化的维度是空间。目前已有的词向量库几乎都是按语言划分的(e.g., Grave et al., 2018; Hamilton et al., 2016)，缺少一个国家内部的州/省/市/县级别的空间细粒度词向量数据，这限制了词向量在探讨文化内差异方面的应用。当然，想获得细粒度的词向量，合适、有效的文本语料

---

[7] 自行训练词向量也存在一定风险，需要特别注意以下问题：(1)由较小规模语料训练得到的词向量可能缺乏准确性和代表性，导致研究结果产生偏差。目前，尚无关于语料规模大小的明确要求，但基本原则是语料越充足，词向量越可靠，尤其要保证语料中有足够多与研究问题有关的词。(2)对于中文词向量的训练，需要事先进行中文分词，以保证后续结果的准确性。如何对中文进行准确分词是一个重要的基础技术问题。研究者可以选用目前较成熟的中文分词工具，如 jieba、HanLP、清华大学 THULAC、北京大学 pkuseg、哈工大 LTP 等。

是必不可少的。遗憾的是, 目前大部分可获取的语料, 包括 Common Crawl、谷歌图书、维基/百度百科、《人民日报》等, 都无法获得详细的地区信息。不过, 一种可行的办法是使用带有用户地区标记的新媒体平台(如新浪微博)或地方性报纸, 为每个省或地区专门训练一个词向量模型。这可能需要花费大量时间和资源来采集数据和训练模型, 但只要形成规模, 将极大促进跨文化心理学研究。

3.5.2　应用动态词向量和语言模型

迄今为止, 虽然词嵌入模型在工业界已经从静态发展到动态, 产生了很多大型预训练语言模型(车万翔 等, 2021), 但是以 Word2Vec、GloVe、FastText 为主的静态词向量仍然是现有心理学研究的主流应用方式。静态词向量将一个词在语料库中的所有上下文信息都压缩到一个向量表示; 然而, 词义可能依语境而变, 更严谨的自然语言分析需要考虑动态词向量(即考虑语境的影响)。

基于动态词向量, Guo 和 Caliskan (2021)提出了语境化词嵌入联系测验(Contextualized Embedding Association Test, CEAT)。通过从研究者感兴趣的语料库随机抽取较大数量的包含目标词和概念词的句子, 然后使用 ELMo、GPT、BERT 模型计算每个词在特定句子中的动态词向量, 可以计算语境化的 WEAT 分数并得到其分布, 进而将不同语境(句子)中的 WEAT 分数视为效应量, 使用随机效应元分析汇总所有语境下的效应量(Guo & Caliskan, 2021)。同样利用动态词向量, 一项最新研究在自然语言中重复验证了大五人格结构, 发现宜人性、外倾性、尽责性是得到较好重复的人格特质维度, 从而为人格心理学的词汇学假设提供了新证据(Cutler & Condon, 2023)。

未来研究不仅要突破静态词向量的局限并利用语境化的动态词向量, 还要尝试打破"向量"这种形式的束缚, 探索直接利用 GPT、BERT 等大型预训练语言模型的可能。本文介绍的词嵌入向量只是自然语言处理的基石而非全貌。未来需要开展大量工作, 发展更优的研究方法和测量工具。

3.5.3　开展跨时间和跨语种的研究

利用词向量的跨时间变化和跨语种差异来考察社会与文化心理在时间上的变迁和空间上的差异是未来两个重要的具体研究方向。

在跨时间研究方面, 以往研究主要使用了预训练好的以年代为单位的 HistWords 词向量库(Hamilton et al., 2016)或自己训练的以年份为单位的词向量数据(e.g., Lawson et al., 2022), 主题涉及社会偏见与刻板印象的变迁、政治意识形态的变迁、文化及其心理含义的变迁等。国内还有学者基于历史语料库和词嵌入技术, 专门开发了用于研究语义演变的 Macroscope 平台(Y. Li et al., 2019)。鉴于社会变迁问题的重要性和前沿性(蔡华俭 等, 2020, 2023; 黄梓航 等, 2018, 2021), 未来研究可以将主题拓展至自我建构、社会动机、群际关系、消费需求、环境态度与行为等方面的变迁, 也可以将时间范围追溯至近代以前, 或将时间粒度细化至月甚至天(取决于能否获得相应的文本语料)。此外, 新近研究发现, 人们越晚习得的、越难进行认知加工的词汇越容易产生历时性的语义演变(Y. Li & Siew, 2022)。因此, 未来还可以继续探究人类对语言的习得和加工如何影响和塑造语义演变。

在跨语种研究方面, 以往研究同样提供了优质、可直接使用的多语种词向量库(Grave et al., 2018), 并从语言的社会心理属性(比如词语是否区分阴阳性)等视角考察了社会偏见等现象(DeFranza et al., 2020), 或从文化相似性、历史相关性、地理邻近性等视角考察了词义表征的跨语言一致性和差异性(Thompson et al., 2020)。未来研究应突破对语言本身的关注, 将多语种词向量数据与国家层面社会生态数据(包括人均 GDP、人口密度、气候条件、农耕方式等)相结合, 并尝试利用计量经济学方法解决因果推断问题, 探索可能的文化心理机制。同时, 研究也要关注语种和国家之间的对应问题, 因为使用同一种语言的国家可能不止一个。

最后, 我们想指出, 虽然词嵌入技术最初源自计算机科学领域对自然语言处理的需要, 对计算编程有一定的要求, 但是近年来, 不同领域的一些前期开拓者已经为运用词嵌入技术开展心理学研究做了大量技术准备, 极大降低了技术门槛(见补充材料 2 和附表 S2)。其中, 本文第一作者基于 R 语言为心理学研究者专门开发了一个免费的词嵌入研究综合工具包: PsychWordVec (Bao, 2022)。运用 PsychWordVec 包, 每一位具有 R 编程基础的心理学研究者都能很快掌握词向量数据的管理与调用、词相似度与 WEAT 等指标的计算和统计分

析、预训练语言模型的调用等, 从而为自己的研究服务。我们期待, 越来越多的国内心理学研究者能及时了解词嵌入这一前沿方法及其在心理学领域的广阔应用前景, 并充分利用 PsychWordVec 等集成化工具包, 将词嵌入真正“嵌入”自己的研究。

参考文献

蔡华俭, 黄梓航, 林莉, 张明杨, 王潇欧, 朱慧珺, … 敬一鸣. (2020). 半个多世纪来中国人的心理与行为变化——心理学视野下的研究. *心理科学进展, 28*(10), 1599−1688.

蔡华俭, 张明杨, 包寒吴霜, 朱慧珺, 杨紫嫣, 程曦, … 王梓西. (2023). 心理学视野下的社会变迁研究: 研究设计与分析方法. *心理科学进展, 31*(2), 159−172.

车万翔, 郭江, 崔一鸣. (2021). *自然语言处理: 基于预训练模型的方法*. 北京: 电子工业出版社.

陈萌, 和志强, 王梦雪. (2021). 词嵌入模型研究综述. *河北省科学院学报, 38*(2), 8−16.

黄梓航, 敬一鸣, 喻丰, 古若雷, 周欣悦, 张建新, 蔡华俭. (2018). 个人主义上升, 集体主义式微? ——全球文化变迁与民众心理变化. *心理科学进展*, *26*(11), 2068−2080.

黄梓航, 王俊秀, 苏展, 敬一鸣, 蔡华俭. (2021). 中国社会转型过程中的心理变化: 社会学视角的研究及其对心理学家的启示. *心理科学进展, 29*(12), 2246−2259.

王垚, 贾宝龙, 杜依宁, 张晗, 陈响. (2022). 基于词向量的多维度正则化 SVM 社交网络抑郁倾向检测方法. *计算机应用与软件, 39*(3), 116−120.

吴胜涛, 杨晨曦, 王世强, 马瑞启, 韩布新. (2020). 正义动机的他人凸显效应: 基于词嵌入联想测验的证据. *科学通报, 65*(19), 2047−2054.

薛栢祥. (2019). *社会媒体语言中外显及内隐社会态度的自动化分析* (硕士学位论文). 天津大学.

杨紫嫣, 刘云芝, 余震坤, 蔡华俭. (2015). 内隐联系测验的应用: 国内外研究现状. *心理科学进展*, *23*(11), 1966−1980.

Agarwal, O., Durupınar, F., Badler, N. I., & Nenkova, A. (2019). Word embeddings (also) encode human personality stereotypes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics* (pp. 205−211), Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1023

Aka, A., & Bhatia, S. (2022). Machine learning models for predicting, understanding, and influencing health perception. *Journal of the Association for Consumer Research, 7*(2), 142−153.

Arseniev-Koehler, A., Cochran, S. D., Mays, V. M., Chang, K.-W., & Foster, J. G. (2022). Integrating topic modeling and word embedding to characterize violent deaths. *Proceedings of the National Academy of Sciences, 119*(10), Article e2108801119.

Bailey, A. H., Williams, A., & Cimpian, A. (2022). Based on billions of words on the internet, PEOPLE = MEN. *Science Advances, 8*(13), Article eabm2463.

Bao, H.-W.-S. (2022). *PsychWordVec: Word embedding research framework for psychological science* [Computer software]. https://CRAN.R-project.org/package=PsychWordVec

Bao, H.-W.-S., Cai, H., & Huang, Z. (2022). Discerning cultural shifts in China? Commentary on Hamamura et al. (2021). *American Psychologist, 77*(6), 786−788.

Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with *SemDis*: An open platform for computing semantic distance. *Behavior Research Methods, 53*, 757−780.

Bengio, Y., Ducharme, R., Vincent, P., & Jauvin, C. (2003). A neural probabilistic language model. *Journal of Machine Learning Research, 3*, 1137−1155.

Bhatia, N., & Bhatia, S. (2021). Changes in gender stereotypes over time: A computational analysis. *Psychology of Women Quarterly, 45*(1), 106−125.

Bhatia, S. (2017a). Associative judgment and vector space semantics. *Psychological Review, 124*(1), 1−20.

Bhatia, S. (2017b). The semantic representation of prejudice and stereotypes. *Cognition, 164*, 46−60.

Bhatia, S. (2019a). Predicting risk perception: New insights from data science. *Management Science, 65*(8), 3800−3823.

Bhatia, S. (2019b). Semantic processes in preferential decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(4), 627−640.

Bhatia, S., Goodwin, G. P., & Walasek, L. (2018). Trait associations for Hillary Clinton and Donald Trump in news media: A computational analysis. *Social Psychological and Personality Science, 9*(2), 123−130.

Bhatia, S., Olivola, C. Y., Bhatia, N., & Ameen, A. (2022). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly, 33*(5), Article 101535.

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences, 29*, 31−36.

Blei, D. M., Ng, A.Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research 3*, 993−1022.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). *Man is to computer programmer as woman is to homemaker? Debiasing word embeddings*. arXiv. https://doi.org/10.48550/arXiv.1607.06520

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017).

Semantics derived automatically from language corpora contain human-like biases. *Science, 356*(6334), 183−186.

Charlesworth, T. E. S., Caliskan, A., & Banaji, M. R. (2022). Historical representations of social groups across 200 years of word embeddings from Google Books. *Proceedings of the National Academy of Sciences, 119*(28), Article e2121798119.

Charlesworth, T. E. S., Yang, V., Mann, T. C., Kurdi, B., & Banaji, M. R. (2021). Gender stereotypes in natural language: Word embeddings show robust consistency across child and adult language corpora of more than 65 million words. *Psychological Science, 32*(2), 218−240.

Chen, H., Yang, C., Zhang, X., Liu, Z., Sun, M., & Jin, J. (2021). From symbols to embeddings: A tale of two representations in computational social science. *Journal of Social Computing, 2*(2), 103−156.

Cutler, A., & Condon, D. M. (2023). Deep lexical hypothesis: Identifying personality structure in natural language. *Journal of Personality and Social Psychology*. Advance online publication. https://doi.org/10.1037/pspp0000443

DeFranza, D., Mishra, H., & Mishra, A. (2020). How language shapes prejudice against women: An examination across 45 world languages. *Journal of Personality and Social Psychology, 119*(1), 7−22.

Du, Y., Fang, Q., & Nguyen, D. (2021). *Assessing the reliability of word embedding gender bias measures*. arXiv. https://doi.org/10.48550/arXiv.2109.04732

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2023). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology, 62*(1), 617−629.

Gandhi, N., Zou, W., Meyer, C., Bhatia, S., & Walasek, L. (2022). Computational methods for predicting and understanding food judgment. *Psychological Science, 33*(4), 579−594.

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences, 115*(16), E3635−E3644.

Grand, G., Blank, I. A., Pereira, F., & Fedorenko, E. (2022). Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour, 6*(7), 975−987.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. arXiv. https://doi.org/10.48550/arXiv.1802.06893

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. K. (1998). Measuring individual differences in implicit cognition: The Implicit Association Test. *Journal of Personality and Social Psychology, 74*(6), 1464−1480.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211−244.

Guo, W., & Caliskan, A. (2021). *Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases*. arXiv. https://doi.org/10.48550/arXiv.2006.03955

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science, 14*(6), 1006−1033.

Hamamura, T., Chen, Z., Chan, C. S., Chen, S. X., & Kobayashi, T. (2021). Individualism with Chinese characteristics? Discerning cultural shifts in China using 50 years of printed texts. *American Psychologist, 76*(6), 888−903.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change*. arXiv. https://doi.org/10.48550/arXiv.1605.09096

Harris, Z. S. (1954). Distributional structure. *Words, 10*(2−3), 146−162.

Heinen, D. J. P., & Johnson, D. R. (2018). Semantic distance: An automated measure of creativity that is novel and appropriate. *Psychology of Aesthetics, Creativity, and the Arts, 12*(2), 144−156.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science, 313*(5786), 504−507.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. *Science, 349*(6245), 261−266.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453−458.

Izzidien, A. (2022). Word vector embeddings hold social ontological relations capable of reflecting meaningful fairness assessments. *AI & Society, 37*, 299−318.

Jackson, J. C., Gelfand, M., De, S., & Fox, A. (2019). The loosening of American culture over 200 years is associated with a creativity−order trade-off. *Nature Human Behaviour, 3*(3), 244−250.

Jackson, J. C., Watts, J., List, J.-M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science, 17*(3), 805−826.

Johnson, D. R., Cuthbert, A. S., & Tynan, M. E. (2021). The neglect of idea diversity in creative idea generation and evaluation. *Psychology of Aesthetics, Creativity, and the Arts, 15*(1), 125−135.

Jonauskaite, D., Sutton, A., Cristianini, N., & Mohr, C.

(2021). English colour terms carry gender and valence biases: A corpus study using word embeddings. *PLoS ONE, 16*(6), Article e0251559.

Joseph, K., & Morgan, J. H. (2020). *When do word embeddings accurately reflect surveys on our beliefs about people?* arXiv. https://doi.org/10.48550/arXiv.2004.12043

Kalyan, K. S., & Sangeetha, S. (2020). SECNLP: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics, 101*, Article 103323.

Karpinski, A., & Steinman, R. B. (2006). The Single Category Implicit Association Test as a measure of implicit social cognition. *Journal of Personality and Social Psychology, 91*(1), 16−32.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review, 84*(5), 905−949.

Kroon, A. C., Trilling, D., & Raats, T. (2021). Guilty by association: Using word embeddings to measure ethnic stereotypes in news coverage. *Journalism & Mass Communication Quarterly, 98*(2), 451−477.

Kurdi, B., Mann, T. C., Charlesworth, T. E. S., & Banaji, M. R. (2019). The relationship between implicit intergroup attitudes and beliefs. *Proceedings of the National Academy of Sciences, 116*(13), 5862−5871.

Kurpicz-Briki, M., & Leoni, T. (2021). A world full of stereotypes? Further investigation on origin and gender bias in multi-lingual word embeddings. *Frontiers in Big Data, 4*, Article 625290.

Lake, B. M., & Murphy, G. L. (2021). Word meaning in minds and machines. *Psychological Review*. Advance online publication. https://doi.org/10.1037/rev0000297

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2), 211−240.

Lawson, M. A., Martin, A. E., Huda, I., & Matz, S. C. (2022). Hiring women into senior leadership positions is associated with a reduction in gender stereotypes in organizational language. *Proceedings of the National Academy of Sciences, 119*(9), Article e2026443119.

Lazer, D., Pentland, A., Adamic, L., Aral, S., Barabási, A.-L., Brewer, D., … van Alstyne, M. (2009). Computational social science. *Science, 323*(5915), 721−723.

Lazer, D. M. J., Pentland, A., Watts, D. J., Aral, S., Athey, S., Contractor, N., … Wagner, C. (2020). Computational social science: Obstacles and opportunities. *Science, 369*(6507), 1060−1062.

Le, T. H., Arcodia, C., Abreu Novais, M., Kralj, A., & Phan, T. C. (2021). Exploring the multi-dimensionality of authenticity in dining experiences using online reviews. *Tourism Management, 85*, Article 104292.

Lee, K., Braithwaite, J., & Atchikpa, M. (2021). Word embedding analysis on colonial history, present issues, and optimism toward the future in Senegal. *Computational and Mathematical Organization Theory, 27*(3), 343−356.

Lenci, A. (2018). Distributional models of word meaning. *Annual Review of Linguistics, 4*, 151−171.

Lewis, M., & Lupyan, G. (2020). Gender stereotypes are reflected in the distributional structure of 25 languages. *Nature Human Behaviour, 4*, 1021−1028.

Li, H.-X., Lu, B., Chen, X., Li, X.-Y., Castellanos, F. X., & Yan, C.-G. (2022). Exploring self-generated thoughts in a resting state with natural language processing. *Behavior Research Methods, 54*, 1725−1743.

Li, M.-H., Li, P.-W., & Rao, L.-L. (2021). Self−other moral bias: Evidence from implicit measures and the Word-Embedding Association Test. *Personality and Individual Differences, 183*, Article 111107.

Li, Y., Engelthaler, T., Siew, C. S. Q., & Hills, T. T. (2019). The Macroscope: A tool for examining the historical structure of language. *Behavior Research Methods, 51*, 1864−1877.

Li, Y., Hills, T., & Hertwig, R. (2020). A brief history of risk. *Cognition, 203*, Article 104344.

Li, Y., & Siew, C. S. Q. (2022). Diachronic semantic change in language is constrained by how people use and learn language. *Memory & Cognition, 50*(6), 1284−1298.

Lin, L., Chen, X., Shen, Y., & Zhang, L. (2020). Towards automatic depression detection: A BiLSTM/1D CNN-based model. *Applied Sciences, 10*(23), Article 8701.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Nicolas, G., Bai, X., & Fiske, S. T. (2021). Comprehensive stereotype content dictionaries using a semi-automated method. *European Journal of Social Psychology, 51*(1), 178−196.

Olson, J. A., Nahas, J., Chmoulevitch, D., Cropper, S. J., & Webb, M. E. (2021). Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences, 118*(25), Article e2022340118.

Prystawski, B., Grant, E., Nematzadeh, A., Lee, S. W. S., Stevenson, S., & Xu, Y. (2020). Tracing the emergence of gendered language in childhood. In S. Denison, M. Mack, Y. Xu, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 1087−1093). Cognitive Science Society. https://cognitivesciencesociety.org/cogsci20/papers/0190/0190.pdf

Rheault, L., & Cochrane, C. (2020). Word embeddings for

the analysis of ideological placement in parliamentary corpora. *Political Analysis, 28*(1), 112−133.
Rice, D., Rhodes, J. H., & Nteta, T. (2019). Racial bias in legal language. *Research and Politics, 6*(2), 1−7.
Richie, R., & Bhatia, S. (2021). Similarity judgment within and across categories: A comprehensive model comparison. *Cognitive Science, 45*(8), Article e13030.
Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology, 5*(1), Article 50.
Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis, 28*(1), 87−111.
Rozado, D., & al-Gharbi, M. (2022). Using word embeddings to probe sentiment associations of politically loaded terms in news and opinion articles from news media outlets. *Journal of Computational Social Science, 5*, 427−448.
Salas-Zárate, R., Alor-Hernández, G., Salas-Zárate, M. d. P., Paredes-Valverde, M. A., Bustos-López, M., & Sánchez-Cervantes, J. L. (2022). Detecting depression signs on social media: A systematic literature review. *Healthcare, 10*(2), Article 291.
Schönemann, P. H. (1966). A generalized solution of the orthogonal Procrustes problem. *Psychometrika, 31*(1), 1−10.
Thompson, B., Roberts, S. G., & Lupyan, G. (2020). Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour, 4*, 1029−1038.
Toney-Wails, A., & Caliskan, A. (2021). *ValNorm quantifies semantics to reveal consistent valence biases across languages and over centuries*. arXiv. https://doi.org/10.48550/arXiv.2006.03950
Tsai, J. L. (2007). Ideal affect: Cultural causes and behavioral consequences. *Perspectives on Psychological Science, 2*(3), 242−259.
Utsumi, A. (2020). Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science, 44*(6), Article e12844.
van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*, 2579−2605.
Wang, B., Xue, B., & Greenwald, A. G. (2019). *Can we derive explicit and implicit bias from corpus?* arXiv. https://doi.org/10.48550/arXiv.1905.13364
Xie, J. Y., Pinto, R. F., Jr., Hirst, G., & Xu, Y. (2019). *Text-based inference of moral sentiment change*. arXiv. https://doi.org/10.48550/arXiv.2001.07209
Xu, H., Zhang, Z., Wu, L., & Wang, C.-J. (2019). The Cinderella Complex: Word embeddings reveal gender stereotypes in movies and books. *PLoS ONE, 14*(11), Article e0225385.
Zhang, Y., Han, K., Worth, R., & Liu, Z. (2020). Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature Communications, 11*, Article 1877.
Zou, W., & Bhatia, S. (2021). Judgment errors in naturalistic numerical estimation. *Cognition, 211*, Article 104647.

Using word embeddings to investigate human psychology: Methods and applications

BAO Han-Wu-Shuang[1,2,3], WANG Zi-Xi[1,2], CHENG Xi[1,2], SU Zhan[1,2], YANG Ying[1,2], ZHANG Guang-Yao[1,2,4], WANG Bo[5], CAI Hua-Jian[1,2]

([1] *CAS Key Laboratory of Behavioral Science, Institute of Psychology, Chinese Academy of Sciences, Beijing* 100101, *China*)
([2] *Department of Psychology, University of Chinese Academy of Sciences, Beijing* 100049, *China*)
([3] *Manchester China Institute, The University of Manchester, Manchester M13 9PL, United Kingdom*)
([4] *State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, and IDG/McGovern Institute for Brain Research, Beijing* 100875, *China*)
([5] *College of Intelligence and Computing, Tianjin University, Tianjin* 300350, *China*)

Abstract: As a fundamental technique in natural language processing (NLP), word embedding quantifies a word as a low-dimensional, dense, and continuous numeric vector (i.e., word vector). Word embeddings can be obtained by using machine learning algorithms such as neural networks to predict the surrounding words given a word or vice versa (Word2Vec and FastText) or by predicting the probability of co-occurrence of multiple words (GloVe) in large-scale text corpora. Theoretically, the dimensions of a word vector reflect

the pattern of how the word can be predicted in contexts; however, they also connote substantial semantic information of the word. Therefore, word embeddings can be used to analyze semantic meanings of text. In recent years, word embeddings have been increasingly applied to study human psychology, including human semantic processing, cognitive judgment, divergent thinking, social biases and stereotypes, and sociocultural changes at the societal or population level. Future research using word embeddings should (1) distinguish between implicit and explicit components of social cognition, (2) train fine-grained word vectors in terms of time and region to facilitate cross-temporal and cross-cultural research, and (3) apply contextualized word embeddings and large pre-trained language models such as GPT and BERT. To enhance the application of word embeddings in psychology, we have developed the R package "PsychWordVec", an integrated word embedding toolkit for researchers to study human psychology in natural language.

Keywords: natural language processing, word embedding, word vector, semantic representation, semantic relatedness, Word Embedding Association Test (WEAT)

附录

补充材料 1　词向量的训练算法和模型

1. 静态词向量: Word2Vec、GloVe、FastText

训练静态词向量的基本策略是：将大规模语料的词共现信息作为机器学习的自监督学习信号(无需人工标注), 利用算法预测词与上下文的共现关系。这种关系既可以是局部语境中的词共现情况, 也可以是全局语境中的词共现矩阵, 两者分别对应了 Word2Vec 和 GloVe 这两种算法。

Word2Vec 是 2013 年由 Google 提出的, 包括两种模型: (1)连续词袋(Continuous Bag-of-Words, CBOW)模型, 在文本中从头至尾依次取同等大小的目标窗口(比如大小为 2 的窗口包括 5 个连续的词), 采用单层神经网络, 根据上下文预测中心词, 得到的词向量为神经网络隐含层权重(一般取输入权重); (2)跳词(Skip-Gram, SG)模型, 同样取一定大小的上下文窗口, 但训练任务是根据中心词预测上下文其他词(Mikolov, Chen et al., 2013)。后者的一个优化方法是负采样, 即负采样跳词(Skip-Gram with Negative Sampling, SGNS)模型: 对于每个训练样本, 按照一定概率生成负样本(不包含当前窗口内词的样本)用于分类训练, 以提高训练效率和语义表征效果(Mikolov, Sutskever et al., 2013)。

GloVe (Global Vectors)是 2014 年由斯坦福大学提出的。与 Word2Vec 仅利用局部上下文不同, GloVe 是预测一定大小的上下文窗口内含有全局统计信息的词−上下文共现矩阵, 并考虑中心词和上下文其他词的位置距离, 将距离更近的词赋予更大权重, 然后进行加权回归, 得到的词向量为回归迭代求解的参数(Pennington et al., 2014)。GloVe 既利用了全局的词共现统计信息, 也考虑了局部上下文语境中词与词之间的位置距离对词共现的影响(Pennington et al., 2014)。因此, GloVe 能更好地反映词与词的共现情况。

此外, 还有一种基于 Word2Vec 的改进算法: FastText, 其基本架构与 Word2Vec 相似, 可以是 CBOW 或 Skip-Gram 模型, 但训练对象不只是单词, 还包括由子词(subword)构成的字符级 *n*-gram, 而且训练时也是预测 *n*-gram 的共现(Bojanowski et al., 2017; Joulin et al., 2016)。FastText 广泛适用于多种语言(Bojanowski et al., 2017)。

附表 S1　目前已有的静态词向量预训练数据库及其使用的语料库

语料库	预训练算法模型		
	Word2Vec (SGNS)	GloVe	FastText
谷歌新闻 (Google News)	√		
谷歌图书 (Google Books)	√(多语种、分年代)		
美式英语历史语料 (COHA)	√(分年代)		
维基百科 (Wikipedia)	√(多语种)	√	√(多语种)
共享网络爬虫 (Common Crawl)		√	√(多语种)
新闻报道千兆语料 (Gigaword)		√	
推特(Twitter)		√	
百度百科	√(汉语)		
新浪微博	√(汉语)		
人民日报	√(汉语)		
搜狗新闻	√(汉语)		
知乎问答	√(汉语)		
四库全书	√(古代汉语)		

注：语料库和预训练算法模型是相互独立的, 表中列出的是目前已有的数据, 默认是英文词向量。读者下载原始的词向量纯文本数据后, 可使用 R 语言 PsychWordVec 包将其转换为 RData 压缩格式, 以便在 R 语言中调用分析。COHA = Corpus of Historical American English.

下载地址：

(1) Google 基于 Google News 训练的词向量库(https://code.google.com/p/word2vec/);

(2)斯坦福大学基于 Google Books 训练的跨年代(1800s~1990s)、多语种(英语、法语、德语、汉语)词向量库 HistWords (Hamilton et al., 2016; https://nlp.stanford.edu/projects/histwords/);

(3)北京师范大学基于百度百科、中文维基百科、人民日报、新闻、微博、知乎、文学作品、四库全书等语料库分别训练的中文词向量库(https://github.com/Embedding/Chinese-Word-Vectors);

(4) GloVe 官方基于 Wikipedia、Common Crawl、Gigaword、Twitter 等大规模语料库分别训练的英文词向量库(Pennington et al., 2014; https://nlp.stanford.edu/projects/glove/);

(5) FastText 官方基于 Wikipedia 和 Common Crawl 大规模语料库训练的全球 157 种语言的词向量库(Grave et al., 2018; https://fasttext.cc/docs/en/crawl-vectors.html)。

目前, 基于这三种静态词向量训练算法, 已经有一系列预训练好的静态词向量数据可供研究者直接下载使用(见附表 S1)。这些静态词向量库

一般使用较大规模的训练语料，因此词汇量较大(几十万至几百万)，能覆盖研究需要的大部分词汇。然而，对于未出现在词向量库中的词，我们无法获取它们的静态词向量。为了根据子词拼接生成整词的向量表示，也为了考虑语境对语义的影响，我们需要利用预训练语言模型来生成动态词向量。

2. 动态词向量和预训练语言模型：ELMo、GPT、BERT

为了完整实现自然语言的理解和生成，需要使模型具备语言编码和解码的能力。这类模型通常被称为预训练语言模型(pre-trained language model)，已不再是简单的词向量训练模型，而是具有语言综合处理能力的复杂模型，参数量更庞大，详细原理可参阅技术文献(车万翔 等, 2021)。预训练语言模型的用途很广，但最基本的用途之一是可以从中提取语境化、动态的词向量，从而解决一词多义(polysemy)问题，使语义的向量化表征更准确。不过，预训练语言模型的词汇量一般较小(一种语言可能只有几万的词汇量)，而且词汇有更多属于子词而非整词。为此，研究者一般可通过对子词向量的叠加来获得词汇表以外(out-of-vocabulary)的整词的向量表示。

目前，Hugging Face 平台(https://huggingface.co/models)已公开存储了万余种预训练语言模型，可供免费下载使用。语言模型的发展极其迅速，从最初的 ELMo (Embeddings from Language Models)动态词向量预训练模型(Peters et al., 2018)，到后来的 GPT (Generative Pre-trained Transformer)生成式预训练模型(Radford et al., 2018)，以及 BERT (Bidirectional Encoder Representations from Transformers)双向编码模型(Devlin et al., 2018)，再到 BERT 的各种衍生模型(如 DistilBERT、ALBERT、RoBERTa、DistilRoBERTa、DeBERTa 等)。本文不再详细介绍，读者可参阅其他资料(车万翔 等, 2021)。

补充材料 2　词向量软件工具简介

附表 S2 总结了 MATLAB、Python 和 R 中与词向量有关的工具包。其中, R 语言 PsychWordVec 包是为心理学专门开发的词嵌入研究综合工具包，推荐读者使用(Bao, 2022)。

附表 S2　词向量相关软件工具的功能简介

编程语言-工具包	可实现的预训练算法	其他功能
MATLAB		
Text Analytics Toolbox	Word2Vec	文本预处理、传统文本分析(词袋模型、潜在语义分析 LSA、主题模型 LDA 等)、文本相似度计算、文本情感分析、词云图绘制等
Python		
gensim 库	Word2Vec、FastText	文本预处理、传统文本分析(词袋模型、潜在语义分析 LSA、主题模型 LDA 等)
fasttext 库	FastText	有监督的文本分类
allennlp 库	ELMo	多种 NLP 下游任务
openai 库	GPT	多种 NLP 下游任务
transformers 库	GPT、BERT 等	预训练语言模型的调用和分析
R		
word2vec 包	Word2Vec	—
text2vec 包	GloVe	分词预处理、潜在语义分析 LSA、主题模型 LDA
fastTextR 包	FastText	—
wordsalad 包	Word2Vec、GloVe、FastText (整合前 3 个包)	—
sweater 包	—	概念联系测量(WEAT、RND 等)的统计分析
PsychWordVec 包	Word2Vec、GloVe、FastText (整合上述 R 包)	词向量数据的统一管理、词向量可视化和 t-SNE 降维、词相似度计算、WEAT (SC-WEAT)与 RND 分析和统计检验、词向量网络分析、不同词嵌入矩阵的正交对齐、预训练语言模型(如 GPT 和 BERT)的调用等

参考文献

车万翔, 郭江, 崔一鸣. (2021). *自然语言处理: 基于预训练模型的方法*. 北京: 电子工业出版社.

Bao, H.-W.-S. (2022). *PsychWordVec: Word embedding research framework for psychological science* [Computer software]. https://CRAN.R-project.org/package=PsychWordVec

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135−146.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of deep bidirectional transformers for language understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. (2013). *Distributed representations of words and phrases and their compositionality*. arXiv. https://doi.org/10.48550/arXiv.1310.4546

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., & Mikolov, T. (2018). *Learning word vectors for 157 languages*. arXiv. https://doi.org/10.48550/arXiv.1802.06893

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change*. arXiv.https://doi.org/10.48550/arXiv. 1605.09096

Joulin, A., Grave, E., Bojanowski, P., & Mikolov, T. (2016). *Bag of tricks for efficient text classification*. arXiv. https://doi.org/10.48550/arXiv.1607.01759

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532−1543), Doha, Qatar. Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. arXiv. https://doi.org/10.48550/arXiv.1802.05365

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Retrieved April 19, 2022 from https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf